

METHOD AND SYSTEM FOR AUTOMATED SESSION RESOURCE CLEAN-UP IN A DISTRIBUTED CLIENT-SERVER ENVIRONMENT

5 Cross Reference to Related Patents

The present invention is related to the following patents which are specifically incorporated herein by reference:

Pending patent application Serial Number 09/409,345 (docket CHA9-1999-0004) filed
10 September 30, 1999 by Cessna et al. entitled "Framework for Dynamic Hierarchical Grouping
and Calculation based on Multidimensional Characteristics" and assigned to the assignee of the
present invention. This patent is sometimes referred to herein as the Framework Patent.

Pending patent application Serial Number 09/491,834 (IBM docket CHA9-99-014) filed
January 26, 2000 by C. Bialik et al. entitled "Method and System for Database Management for
15 Supply Chain Management" and assigned to the assignee of the present invention. This patent is
sometimes referred to herein as the Database Patent.

Patent application Serial Number (to be assigned for IBM docket CHA9-99-013) filed
concurrently by the inventor of the present document, Iwao Hatanaka, and entitled "Method and
System for Incorporating Legacy Applications into a Distributed Data Processing System" and
20 assigned to the assignee of the present invention. This patent is sometimes called the Legacy
Application Patent.

Issued United States patent number 6,021,493 of Daryl C. Cromer et al. entitled "System
and Method for Detecting When a Computer System is Removed from a Network" issued on
February 1, 2000 and assigned to the assignee of the present invention. This patent is sometimes
25 referred to herein as the Heartbeat Patent and is useful in detecting whether a client is attached to
a server.

Background of the Invention

Field of the Invention

5 The present invention is an improved system and method for automatically managing session resources in a distributed network of processors, such as a client-server environment, where the invention has the particular advantage of automatically releasing those resources allocated to a session when the session ends, whether through a normal ending or through an abnormal ending. More particularly, the present invention includes a session management
10 framework which can be applied to release session resources when the session ends abnormally, e.g., through the halting of an application or the loss of a connection between the server and the client.

Background Art

15 In a client-server environment, a local terminal (sometimes referred to as a client) is connected to a server for the purpose of processing information in a distributed environment. Frequently, the client is itself a data processing system which communicates with a server which is generally a data processing system with increased resources, including applications and data which are not available at the client application. Such a system is described in some detail in the Framework Patent referenced above. In a client-server environment, resources may be centrally
20 managed at the server as opposed to being disparately managed at each individual client. In some cases, the client does not have the capability of managing or maintaining large resources.

 The client is frequently located at a distance from the server and communicates with the server using telecommunication facilities, including hardware and software operating over phone service such as might be provided using telephone lines, either alone or in combination with

other communication systems such as satellite or microwave communications. A series of communications occur between a client and its server (for example, to execute an application on the server using data supplied by the client and report the results of the application back to the client) is sometimes referred to as a session, with a session including a plurality of

5 communications between the server and the client. In any event, there are frequently several different links in the communications chain, and when one of the links fails to operate, the communications channel is disrupted and the session is terminated.

While the session was in existence, various resources at the server are dedicated or reserved for the use of the particular client which requests use of those resources. So, in a
10 supply chain application, a variety of storage units associated with the server may be used by a client during a session and various applications and databases may be dedicated to the client and its session, often to the preclusion of using those same resources for other clients while a session with the one client is in progress. Such a preclusion is understandable, particularly when an application may be changing the application or the database, so access during a change by
15 another application may provide the wrong execution or the wrong data.

A session with a client “ties up” resources generally (memory used for one application cannot be used at the same time for another application) and for some specific reasons (a client which is using a database typically marks the database so that another client cannot simultaneously use the database and change the information stored in the database while the
20 other client is using the database, for example).

Since the resources are limited and other clients may want to use the same resources, it is advantageous to release the resources as soon as the resources are not needed, and the normal termination of a session (e.g., the completion of execution of a program) typically provides a

release of the resources which have been used for the session as a part of the normal ending of the session.

But, when a session is abnormally terminated, it does not go through the normal ending or winding down process which releases the resources. In fact, many of the events which
5 contribute to the abnormal termination of a session result from a total lack of communications with a client, perhaps because the connection between the client and the server is no longer functional. This is becoming more of a problem when the communication is over the public Internet or a virtual private networks, where a large number of users are connected through paths which are constantly changing as the network evolves, and the session depends on the continuing
10 availability of a path between the client and the server.

The Legacy Application Patent describes an approach to allow use of legacy applications in a distributed processing environment, allowing legacy applications which were not designed to be utilized in a distributed processing system to be used in such a system. Such a system inherently requires that resources which are being used in a distributed data processing system be
15 committed to the use and be released once the processing has ended.

Several approaches have been suggested for determining when a session is no longer active. One of these involves polling, or making sure that the client and the server remain active by periodically issuing an inquiry from the one to the other with an answer back if the connection is still in place. This involves setting up some kind of periodic inquiry system and
20 keeping track of when an inquiry is due for each of the clients, an exercise which requires resources and does not necessarily provide a prompt notice that a client has been dropped by the network -- that is, without a lot of repeated polling of each client every short interval, the server does not know which clients remain attached and which clients are no longer attached. But,

polling requires continuing use of resource and suggests that polling ought to be done at lengthy intervals to reduce the use of network resources, but the longer the interval, the longer resources may be dedicated to serve a session which no longer exists.

A prior art system for determine whether a resource is attached sometimes uses a “heart beat” technique for determining whether the resource remains attached. But, in such a system a ping is sent out addressed to the remote user and the absence of a response is taken to mean that the resource is not attached, when, in fact, the ping or its response may have been misdirected or lost in the system without the resource actually being disconnected. Another disadvantage of polling is that message traffic is increased for each client which is added to the system. Also, there is the lack of an unequivocal indication that a resource is no longer needed or that a client is no longer connected.

Accordingly, the prior art systems have undesirable disadvantages and limitations.

Summary of the Invention

The present invention overcomes the limitations and disadvantages of the prior art systems by providing a system and method for releasing resources dedicated to a session promptly, even when the session ends abnormally and without a termination message.

The present invention has the advantage that it is simple and easy to implement to allow for the release of resources held for a client when the client is no longer connected to the server.

The present invention allows for the prompt reallocation of resources from a client to an other client when the first client is no longer using the resource without polling or a periodic inquiry of the connected status of each of the clients using resources of a given server.

The present invention involves setting up a resource manager for each session and logging the use of resources associated with that session. Then, when the session is no longer active -- for whatever reason, including normal disconnection or a lost connection, the resource manager consults the listing of resources associated with that session and releases the resources
5 for use, allowing use with other sessions.

Using the resource manager for normal and abnormal session terminations means that it is not necessary to have two different types of session terminations, one for normal terminations and a different one for an abnormal termination.

The present system also allows for a table which identifies which resource is associated
10 with which user.

The present application is suited for use in a system such as are described in the Legacy Application Patent. The use of a distributed data processing solution means that different processors may have reserved resources such as applications which need to be resolved when a session ends.

15 A system such as the Heartbeat Patent may be used to determine whether a client is attached to the server at any given time. By periodically querying the clients, it is possible to determine whether the client is still coupled to the server or if the connection has been lost for some reason. The Heartbeat Patent is one way (but certainly not the only way) to determine whether the client is still coupled to the server and capable of communicating. If the Heartbeat
20 Patent detects that a given client is no longer attached, it can signal the server to allow release of the resources associated with the client.

Other objects and advantages of the present invention will be apparent to those skilled in the relevant art in view of the following description of the preferred embodiment, taken together with the accompanying drawings and the appended claims.

5

Brief Description of the Drawings

10

The present invention is an improved system and method for resource cleanup, an embodiment of which is illustrated with reference to the accompanying drawings in which:

Fig. 1 depicts a communications system representative of the preferred embodiment of the present invention;

15

Fig. 2, consisting of Fig. 2A and Fig. 2B, are flow diagrams of the preferred embodiment of the present invention; and

Fig. 3, consisting of Fig. 3A and Fig. 3B, are diagrams illustrating resource tables useful in practicing the present invention.

20

Detailed Description of the Preferred Embodiment

In the following description of the preferred embodiment, the best implementation of practicing the invention presently known to the inventors will be described with some particularity. However, this description is intended as a broad, general teaching of the concepts of the present invention in a specific embodiment but is not intended to be limiting the present

25

invention to that as shown in this embodiment, especially since those skilled in the relevant art will recognize many variations and changes to the specific structure and operation shown and described with respect to these figures.

Fig. 1 illustrates a communications system of the type used in the present invention. In

5 this Fig. 1, a first client (CLIENT1) 100 is connected to a first server (SERVER1) 110 through a network 120. Additional clients (CLIENT2, CLIENT3, CLIENT4) 131, 132, 133, respectively are shown also connected to the first server 110 through the network 120 and additional servers (SERVER2, SERVER3 and SERVER4) 141, 142, 143, respectively are also shown connected to the network 120. While this is a simplistic view of a network in which a plurality of servers are
10 connected to serve a plurality of clients, it will allow discussion of the problems with such an arrangement and an understanding of the present invention and its advantages. The first client 100 may involve an application which uses a resource at the first server 110 (for example, an application APPLN1 referred to by the reference numeral 111) and a resource at the second server 141 (for example, a database DB referred to by the reference numeral 151) and store the
15 result in a file 152 maintained on the third server 142 (the file 152 might be a file with pro forma income and profit projections), all of which data processing is accomplished through the communications network 120 which connects the client 110 with the servers 110, 141 and 142. Meanwhile, the second client 131 may wish to use resources at the first server 110, the second server 141 and a fourth server 143. If the second client 131 is using different resources at the
20 servers from the other clients at any given time, then there is no problem. If, however, the first client 100 is using the particular application APPLN1 111 at the first server 110, then the second client may not be permitted to use the application APPLN1 111 at that same time, but would be permitted to use an application APPLN2 112 which is also at the first server.

The present invention leverages the fact that each client session with a server is associated with a single file descriptor in the server during a client connection to the server. All communications from and to that client takes place through that file descriptor. Through a callback program associated with that file descriptor, client termination events can be captured to trigger desired system processing at precisely the time that the client disconnects from the server. This functionality allows for automatic session clean-up by detecting client termination and then freeing up corresponding resources being held on the server for the terminated client session.

Fig. 2 illustrates in flow diagram form the logic of the present invention showing aspects of the present invention. Fig. 2 consists of Fig. 2A and Fig. 2B. Fig. 2A shows logic for the determination of whether a resource is available and assigning the resource to a particular requesting client while Fig. 2B shows logic for determining whether to release a resource and the steps taken to release that resource and allow for further use of the resource by other clients.

Fig. 2A illustrates the process of a client using resources at a server as was described in connection with Fig. 1. The process starts at block 202 and at block 204 a request is received by the server for resources associated with that server, resources which may be use of an application, access to a database stored on the server or simply to a block of memory, for example, as a temporary storage for an application. While the server may have a large number of resources and many of these resources are not unique (one block of empty memory may be similar to the next), others of the resources are unique (the server may have a single copy of an application or a database) and the resources are limited (the server might well run out of memory if the memory were not released and reused by a second client after the first has completed its processing). Based on the request received at the block 204 for resources, at block 206 the server

determines whether the resource is available to the requesting client. Such availability is determined in connection with resource listings such as Fig. 3, particularly Fig. 3B which identifies each resource as being available or being used by a named client. If the client is requesting use of a database already in use by another client or if the memory requested is not available, then the request is denied at block 208 with an appropriate message ('resource in use; try again later' or "inadequate memory presently available; try elsewhere or try again later"). If, on the other hand, the resource is available for the client, then at block 210, access is granted and the resource is logged (see Fig. 3 and the associated text for a discussion of the logging process to include identification of which resources are available and which are used by which clients) as assigned to the requesting client. In any event, following the disposition of a request for resources, either by granting it at block 210 or denying it at block 208, control returns to the starting area where the next request can be processed by the block 204.

Fig. 2B illustrates the process for releasing a resource which has been assigned to a client and for which the client no longer has a need for the resource. Such a release may be because the program using the resource has run its course and terminated successfully or because something unnatural has occurred, like the client has become disconnected from the server -- i.e., either the server 110 or the client 100 is no longer connected to the network 120 or the client 100 is no longer operational. While a normal termination of an application program may issue the explicit command to release the resources that the application has been using, the program may abort or otherwise not issue such a command.

The process of Fig. 2B is as follows: starting from block 220, at block 222 the question is asked whether a client has specifically released a resource. If not, then at block 224, it is determined whether the client remains attached to the network. This determination can be made

through any of a number of conventional approaches, such as “pinging” the client or by determining a heartbeat of the client using the Heartbeat Patent referenced above. If the client is present, then control passes to an optional set of time determinations which serve to limit the time that a resource can be used -- either with activity or without activity. Associated with the

5 resource (e.g., an application, database or memory) and/or the client are allowable time intervals. For example, a client may use a first application for 30 minutes but will be considered inactive if no activity occurs within a 15 minute time period. Thus, at block 226 the amount of time a resource has been used will be compared with an allowable time for such use (if any has been set) by comparing the present time with the beginning time which was stored in column
10 308 of Fig. 3A to determine the amount of time the resource has been in use. If the time that the resource has been used does not exceed the limit, then the amount of inactive time is compared at block 228. That is, the period since the last use (in column 310 of Fig. 3) to the present is compared with a threshold (if set) to determine whether the resource has been held without activity longer than a preset period of time. If the client released the resource (at block 222), the
15 client is not attached (at block 224), the time of use (block 226) or the time of inactivity (block 228) exceed the set limits, then the resource is released at block 230 with the entry in the table of resources being used (Fig. 3A) erased at block 232 and the resource marked as available in the listing of Fig. 3B at block 234. Control then returns to the start for the next resource action.

Fig. 3 shows resource tables useful in practicing the present invention. In Fig. 3A, a first
20 table 300 depicts in list form the resources currently being used and the client using each of the resources. Although only a portion of this table 300 is shown to illustrate the principles of the present invention, the table could be as large as necessary to contain data about all the clients using the server and the resources that each of the clients is currently using. The table includes a

first column 302 which lists the resource being used, a second column 304 listing the client using the resource, a third column 306 indicating the type of access (whether it is read only or read/write), a fourth column 308 indicating the time which the resource was first accessed and a fifth column 310 indicating the time that the resource was last used. Use of the fourth column 308 with the beginning time allows for a time limit to be set for release of the resource after a fixed amount of time and the fifth column 310 (last use) allows for a time limit to be set that releases a resource if it has not been used within a fixed period of time. That is, the resource could be released after x minutes of use (based on a comparison of the current time with the start time stored in column 308) or after y minutes of nonuse (based on a comparison of the current time with the time in column 310). The times allowed (x minutes of use, y minutes of nonuse) are subject to system constraints and may be adjusted based on the type of use and whether concurrent uses are permitted. In some situations, a read-only access of a resource may not preclude others' use of the same resource and one client might be permitted to continue to use such a resource on a non-exclusive basis than would be permitted if the resource were being used on an exclusive basis. An optional sixth column 312 provides the time of the last indication that the client is connected, a time which may be provided by receiving a request from the client or from a return "ping" of the client as discussed elsewhere.

In Fig. 3B, a listing of the resources is provided and an associated status for each resource -- whether the resource is "free" for use by a client or if it is currently committed to a client and not available. This Fig. 3B lists each of the resources along with its status, by resource. So, Fig. 3B includes a left column 330 which lists the resource and a right column 332 which either lists the resource as being used by a named client or being available. This, for the simple example of Fig. 1, APPLN1 is shown in block 334 as a resource and in block 344, it is being used by client

110. APPLN2 is listed in block 336 as a resource and in block 346, it is being used by client 131. APPLN3 is listed in block 338 as being available in block 348. Similarly blocks of memory and other resources such as the database DB can be assigned to a particular client, and at the end of the use by that client, release by removing the entry in the columns of Fig. 3.

5 The present invention can be realized in hardware, software, or a combination of hardware and software. A data processing tool according to the present invention can be realized in a centralized fashion in one computer system, or in a distributed fashion where different elements are spread across several interconnected computer systems. Any kind of computer system - or other apparatus adapted for carrying out the methods described herein - is suited. A
10 typical combination of hardware and software could be a general purpose computer system with a computer program that, when being loaded and executed, controls the computer system such that it carries out the methods described herein. The present invention can also be embedded in a computer program product, which comprises all the features enabling the implementation of the methods described herein, and which - when loaded in a computer system - is able to carry out
15 these methods.

“Computer program means” or “computer program” in the present context mean any expression, in any language, code or notation, of a set of instructions intended to cause a system having an information processing capability to perform a particular function either directly or after either or both of the following a) conversion to another language, code or notation; b)
20 reproduction in a different material form.

While the present invention is described in the context of an apparatus and a method of providing resource management, the present invention may be implemented in the form of a

service where collecting, maintaining and processing of information is located apart from the server and information is communicated as needed to the server.

Of course, many modifications of the present invention will be apparent to those skilled in the relevant art in view of the foregoing description of the preferred embodiment, taken
5 together with the accompanying drawings. For example, the system for recognizing that the session between the client and the server no longer exists may be determined in any manner and is not limited to that disclosed in the foregoing material. Additionally, the location and type of information maintained about a session may be modified to suit the application and need not be the listing of resources associated with each client as disclosed. Such information may be stored
10 in connection with each resource being used rather than in a central location, although there are advantages to having the information located centrally in that a central location makes it easier and quicker to release and reuse the resource again. Additionally, certain features of the present invention may be useful without the corresponding use of other features without departing from the spirit of the present invention. For example, a client may be using several resources
15 associated with different applications and one application may end (so the resources associated with that application should be released) or the entire connection may terminate (so all applications terminate). Further, the system of Fig. 3B arranges data on resource use including the same data as Fig. 3A, and the two could be combined, if desired, using a single database to show what resources are in use and what clients are using the resource. Accordingly, the
20 foregoing description of the preferred embodiment should be considered as merely illustrative of the principles of the present invention and not in limitation thereof.